

# PATENT COOPERATION TREATY

PCT

## NOTIFICATION CONCERNING SUBMISSION OR TRANSMITTAL OF PRIORITY DOCUMENT

(PCT Administrative Instructions, Section 411)

From the INTERNATIONAL BUREAU


To:

KOIKE, Akira  
No.11 Mori Bldg., 6-4, Toranomón 2-  
chome  
Minato-ku, Tokyo 105-0001  
JAPON

Date of mailing (day/month/year) 02 February 2000 (02.02.00)	
Applicant's or agent's file reference SK00PCT7	<b>IMPORTANT NOTIFICATION</b>
International application No. PCT/JP00/00203	International filing date (day/month/year) 18 January 2000 (18.01.00)
International publication date (day/month/year) Not yet published	Priority date (day/month/year) 21 January 1999 (21.01.99)
Applicant SONY CORPORATION et al	

- The applicant is hereby notified of the date of receipt (except where the letters "NR" appear in the right-hand column) by the International Bureau of the priority document(s) relating to the earlier application(s) indicated below. Unless otherwise indicated by an asterisk appearing next to a date of receipt, or by the letters "NR", in the right-hand column, the priority document concerned was submitted or transmitted to the International Bureau in compliance with Rule 17.1(a) or (b).
- This updates and replaces any previously issued notification concerning submission or transmittal of priority documents.
- An asterisk(\*) appearing next to a date of receipt, in the right-hand column, denotes a priority document submitted or transmitted to the International Bureau but not in compliance with Rule 17.1(a) or (b). In such a case, **the attention of the applicant is directed** to Rule 17.1(c) which provides that no designated Office may disregard the priority claim concerned before giving the applicant an opportunity, upon entry into the national phase, to furnish the priority document within a time limit which is reasonable under the circumstances.
- The letters "NR" appearing in the right-hand column denote a priority document which was not received by the International Bureau or which the applicant did not request the receiving Office to prepare and transmit to the International Bureau, as provided by Rule 17.1(a) or (b), respectively. In such a case, **the attention of the applicant is directed** to Rule 17.1(c) which provides that no designated Office may disregard the priority claim concerned before giving the applicant an opportunity, upon entry into the national phase, to furnish the priority document within a time limit which is reasonable under the circumstances.

<u>Priority date</u>	<u>Priority application No.</u>	<u>Country or regional Office or PCT receiving Office</u>	<u>Date of receipt of priority document</u>
21 Janu 1999 (21.01.99)	11/13307	JP	28 Janu 2000 (28.01.00)

The International Bureau of WIPO 34, chemin des Colombettes 1211 Geneva 20, Switzerland  Facsimile No. (41-22) 740.14.35	Authorized officer  Y. KUWAHARA   Telephone No. (41-22) 338.83.38
--	--

This Page Blank (uspic)

日本国特許庁

PATENT OFFICE  
JAPANESE GOVERNMENT

18.01.00	
REC'D 28 JAN 2000	
WIPO	PCT

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日  
Date of Application:

1999年 1月21日

出願番号  
Application Number:

平成11年特許願第013307号

出願人  
Applicant(s):

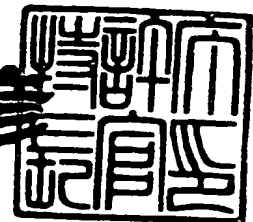
ソニー株式会社

PRIORITY  
DOCUMENT  
SUBMITTED OR TRANSMITTED IN  
COMPLIANCE WITH RULE 17.1(a) OR (b)

1999年12月 3日

特許庁長官  
Commissioner,  
Patent Office

近藤隆彦



出証番号 出証特平11-3084598

【書類名】 特許願

【整理番号】 9801108806

【あて先】 特許庁長官殿

【国際特許分類】 G06F 7/00

【発明者】

    【住所又は居所】 東京都品川区北品川 6 丁目 7 番 3 5 号 ソニー株式会社  
内

    【氏名】 長尾 確

【特許出願人】

    【識別番号】 000002185

    【氏名又は名称】 ソニー株式会社

    【代表者】 出井 伸之

【代理人】

    【識別番号】 100067736

    【弁理士】

    【氏名又は名称】 小池 晃

【選任した代理人】

    【識別番号】 100086335

    【弁理士】

    【氏名又は名称】 田村 榮一

【選任した代理人】

    【識別番号】 100096677

    【弁理士】

    【氏名又は名称】 伊賀 誠司

【手数料の表示】

    【予納台帳番号】 019530

    【納付金額】 21,000円

【提出物件の目録】

    【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1

【包括委任状番号】 9707387

【プルーフの要否】 要

【書類名】 明細書

【発明の名称】 文書処理方法および装置ならびに記録媒体

【特許請求の範囲】

【請求項 1】 複数の要素から構成される内部構造を有する文書进行处理する文書処理方法において、

上記文書についてその文書の特徴を表す特徴情報を抽出する特徴情報抽出工程と、

分類モデルを構成する複数の分類項目について、上記特徴情報抽出工程で抽出した文書の特徴情報と上記分類項目毎の特徴情報との関連度に応じて、各文書を上記分類項目に分類する文書分類工程と

を有することを特徴とする文書処理方法。

【請求項 2】 複数の文書を受信する受信工程を有し、

上記特徴情報抽出工程は、上記受信工程で受信した各文書についてその文書の特徴を表す特徴情報を抽出すること

を特徴とする請求項 1 記載の文書処理方法。

【請求項 3】 上記特徴情報抽出工程は、上記文書の内部構造に基づいて各要素に重みを付与し、この重みが所定値より大きい要素を抽出すること

を特徴とする請求項 1 記載の文書処理方法。

【請求項 4】 上記文書分類工程は、複数の文書を分類項目に分類した分類操作に基づいて作成された分類モデルを用いて文書を分類すること

を特徴とする請求項 1 記載の文書処理方法。

【請求項 5】 上記文書分類工程は、上記特徴情報抽出工程で抽出した文書の特徴情報と上記分類項目の特徴情報との関連度を計算し、上記関連度が閾値を超えると、上記関連度が最大となる分類項目に文書を分類すること

を特徴とする請求項 1 記載の文書処理方法。

【請求項 6】 上記特徴情報抽出工程で抽出した文書の特徴情報と上記分類項目の特徴情報の関連度は、上記文書の特徴情報に含まれる固有名詞と上記分類項目の特徴情報に含まれる固有名詞とに共通する固有名詞の数と、上記文書の特徴情報に含まれる固有名詞以外の語義の上記分類項目に含まれる固有名詞以外の語

義に対する語義の関連度の総和との線形結合であること

を特徴とする請求項 5 記載の文書処理方法。

【請求項 7】 上記語義の関連度は、語義の間の参照関係の構造に基づいて各語義に重みを付与し、一の語義に付与された重みを独立変数とするとともに他の語義に付与された重みを従属変数とし、上記他の語義に付与された重みの差分を上記一の語義に付与された重みの差分で除した商を上記一の語義と上記他の語義の関連度とすることにより得られたものであること

を特徴とする請求項 6 記載の文書処理方法。

【請求項 8】 上記語義の間の参照関係は、各語義について他の語義を参照する辞書を用いて作成されたこと

を特徴とする請求項 7 記載の文書処理方法。

【請求項 9】 複数の要素から構成される内部構造を有する文書を処理する文書処理方法において、

語義の間の参照関係の構造に基づいて各要素に重みを付与する重み付与工程と

上記重み付与工程で一語義に付与された重みを独立変数とするとともに他の語義に付与された重みを従属変数とし、上記他の語義に付与された重みの差分を上記一の語義に付与された重みの差分で除した商を上記一の語義と上記他の語義の関連度として演算する関連度演算工程と

を有することを特徴とする文書処理方法。

【請求項 10】 上記重み付与工程は、各語義について他の語義を参照する辞書を用いて語義の参照関係を組織した参照関係の構造に基づいて重みを付与すること

を特徴とする請求項 9 記載の文書処理方法。

【請求項 11】 上記辞書の各語義にはその属性を示す属性情報が付与され、上記参照関係の構造は上記属性情報に基づいて組織されること

を特徴とする請求項 10 記載の文書処理方法。

【請求項 12】 複数の要素から構成される内部構造を有する文書を処理する文書処理装置において、

上記文書についてその文書の特徴を表す特徴情報を抽出する特徴情報抽出手段と、

分類モデルを構成する複数の分類項目について、上記特徴情報抽出手段で抽出した文書の特徴情報と上記分類項目毎の特徴情報との関連度に応じて、各文書を上記分類項目に分類する文書分類手段と

を有することを特徴とする文書処理装置。

【請求項 13】 複数の要素から構成される内部構造を有する文書进行处理する文書処理装置において、

語義の間の参照関係の構造に基づいて各要素に重みを付与する重み付与手段と

上記重み付与手段で一語義に付与された重みを独立変数とするとともに他の語義に付与された重みを従属変数とし、上記他の語義に付与された重みの差分を上記一の語義に付与された重みの差分で除した商を上記一の語義と上記他の語義の関連度として演算する関連度演算手段と

を有することを特徴とする文書処理装置。

【請求項 14】 複数の要素から構成される内部構造を有する文書进行处理する文書処理プログラムが記録された記録媒体において、上記文書処理プログラムは

上記文書についてその文書の特徴を表す特徴情報を抽出する特徴情報抽出処理と、

分類モデルを構成する複数の分類項目について、上記特徴情報抽出処理で抽出した文書の特徴情報と上記分類項目毎の特徴情報との関連度に応じて、各文書を上記分類項目に分類する文書分類処理と

を有することを特徴とする記録媒体。

【請求項 15】 複数の要素から構成される内部構造を有する文書进行处理する文書処理プログラムが記録された記録媒体において、上記文書処理プログラムは

語義の間の参照関係の構造に基づいて各要素に重みを付与する重み付与処理と



上記重み付与処理で一の語義に付与された重みを独立変数とするとともに他の語義に付与された重みを従属変数とし、上記他の語義に付与された重みの差分を上記一の語義に付与された重みの差分で除した商を上記一の語義と上記他の語義の関連度として演算する関連度演算処理と

を有することを特徴とする記録媒体。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明は、要素について内部構造を付与された文書进行处理する文書処理方法および装置ならびに上記文書进行处理するプログラムを記録された記録媒体に関し、詳しくは、文書に含まれる語義の関連度に基づいて文書を分類するような文書処理方法および装置、ならびに文書に含まれる語義の関連度に基づいて文書を分類するような文書処理のプログラムが記録されてなる記録媒体に関する。

【0002】

【従来の技術】

従来、インターネットにおいて、ウィンドウ形式でハイパーテキスト型情報を提供するアプリケーションサービスとしてWWW (World Wide Web) が提供されている。

【0003】

WWWは、文書の作成、公開または共有化の文書処理を実行し、新しいスタイルの文書の在り方を示したシステムである。しかし、文書の実際上の利用の観点からは、文書の内容に基づいた文書の分類や要約といった、WWWを越える高度な文書処理が求められている。このような高度な文書処理には、文書の内容の機械的な処理が不可欠である。

【0004】

しかしながら、文書の内容の機械的な処理は、以下のような理由から依然として困難である。第1に、ハイパーテキストを記述する言語であるHTML (Hyper Text Markup Language) は、文書の表現については規定するが、文書の内容についてはほとんど規定しない。第2に、文書間に構成されたハイパーテキストの

ネットワークは、文書の読者にとって文書の内容を理解するために必ずしも利用しやすいものではない。第3に、一般に文章の著作者は読者の便宜を念頭に置かずに著作するが、文書の読者の便宜が著作者の便宜と調整されることはない。

【0005】

このように、WWWは新しい文書の在り方を示した革新的なシステムであるが、文書を機械的に処理しないために、高度な文書処理を行うことができなかった。換言すると、高度な文書処理を実行するためには、文書を機械的に処理することが必要となる。

【0006】

そこで、文書の機械的な処理を目標として、文書の機械的な処理を支援するシステムが自然言語研究の成果に基づいて開発されている。自然言語研究による文書処理の最初のステップとして、文書の著作者等による文書の内部構造についての属性情報、いわゆるタグの付与を前提とした、文書に付与されたタグを利用する機械的な文書処理が提案されている。

【0007】

【発明が解決しようとする課題】

ところで、近年のコンピュータの普及や、ネットワーク化の進展に伴い、文章処理や、文書の内容に依存した索引などで、テキスト文書の作成、ラベル付け、変更などを行う文書処理の高機能化が求められている。このような高機能な文書処理を実現するためには、文書内における各語義の関連度に基づいた文書処理が必要となる。

【0008】

本発明は、上述の実情に鑑みて提案されるものであって、文書内における語義の関連度に基づいた文書処理を行うような文書処理方法および装置、ならびに文書内における関連度に基づいた文書処理のプログラムが記録されてなる記録媒体を提供することを目的とする。

【0009】

【課題を解決するための手段】

上述の課題を解決するために、本発明に係る文書処理方法は、複数の要素から

構成される内部構造を有する文書処理する文書処理方法において、上記文書についてその文書の特徴を表す特徴情報を抽出する特徴情報抽出工程と、分類モデルを構成する複数の分類項目について、上記特徴情報抽出工程で抽出した文書の特徴情報と上記分類項目毎の特徴情報との関連度に応じて、各文書を上記分類項目に分類する文書分類工程とを有するものである。

## 【0010】

また、本発明に係る文書処理方法は、複数の要素から構成される内部構造を有する文書処理する文書処理方法において、語義の間の参照関係の構造に基づいて各要素に重みを付与する重み付与工程と、上記重み付与工程で一の語義に付与された重みを独立変数とするとともに他の語義に付与された重みを従属変数とし、上記他の語義に付与された重みの差分を上記一の語義に付与された重みの差分で除した商を上記一の語義と上記他の語義の関連度として演算する関連度演算工程とを有するものである。

## 【0011】

本発明に係る文書処理装置は、複数の要素から構成される内部構造を有する文書処理する文書処理装置において、上記文書についてその文書の特徴を表す特徴情報を抽出する特徴情報抽出手段と、分類モデルを構成する複数の分類項目について、上記特徴情報抽出手段で抽出した文書の特徴情報と上記分類項目毎の特徴情報との関連度に応じて、各文書を上記分類項目に分類する文書分類手段とを有するものである。

## 【0012】

また、本発明に係る文書処理装置は、複数の要素から構成される内部構造を有する文書処理する文書処理装置において、語義の間の参照関係の構造に基づいて各要素に重みを付与する重み付与手段と、上記重み付与手段で一の語義に付与された重みを独立変数とするとともに他の語義に付与された重みを従属変数とし、上記他の語義に付与された重みの差分を上記一の語義に付与された重みの差分で除した商を上記一の語義と上記他の語義の関連度として演算する関連度演算手段とを有するものである。

## 【0013】

本発明に係る記録媒体は、複数の要素から構成される内部構造を有する文書を処理する文書処理プログラムが記録された記録媒体において、上記プログラムは、上記文書についてその文書の特徴を表す特徴情報を抽出する特徴情報抽出処理と、分類モデルを構成する複数の分類項目について、上記特徴情報抽出処理で抽出した文書の特徴情報と上記分類項目毎の特徴情報との関連度に応じて、各文書を上記分類項目に分類する文書分類処理とを有するものである。

## 【0014】

また、本発明に係る記録媒体は、複数の要素から構成される内部構造を有する文書を処理する文書処理プログラムが記録された記録媒体において、語義の間の参照関係の構造に基づいて各要素に重みを付与する重み付与処理と、上記重み付与処理で一語義に付与された重みを独立変数とするとともに他の語義に付与された重みを従属変数とし、上記他の語義に付与された重みの差分を上記一の語義に付与された重みの差分で除した商を上記一の語義と上記他の語義の関連度として演算する関連度演算処理とを有するものである。

## 【0015】

## 【発明の実施の形態】

以下、図面を参照して、本発明に係る文書処理方法および装置ならびに記録媒体の実施の形態について説明する。

## 【0016】

本発明の実施の形態としての文書処理装置は、図1に示すように、制御部11およびインターフェース12を備える本体10と、ユーザからの入力を受け付けて本体10に送る入力部20と、外部からの信号を受信して本体10に送る受信部21と、本体10からの出力を表示する表示部30と、記録媒体32に対して情報を記録／再生する記録／再生部312とを有している。

## 【0017】

本体10は、制御部11およびインターフェース12を有し、この文書処理装置の主要な部分を構成している。制御部11は、この文書処理装置における処理を集中して実行するCPU13と、揮発性のメモリであるRAM14と、不揮発

性のメモリであるROM15とを有している。CPU13は、たとえばROM14に記録された手順にしたがって、必要な場合にはデータを一時的にRAM14に格納して、プログラムを実行するための制御を行う。インターフェース12には、入力部20、受信部21および表示部30が接続される。インターフェース12は、制御部11からの制御の下に、入力部20および受信部21からのデータの入力、表示部30へのデータの送信について、データを送信するタイミングを調整したり、データの形式を変換したりする。

## 【0018】

入力部20は、この文書処理装置に対するユーザの入力を受け付ける部分である。この入力部20は、たとえばキーボードやマウスにより構成される。ユーザは、この入力部20を用い、キーボードによりキーワードを入力したり、マウスにより表示部30に表示されている文書のエレメントを選択して入力したりすることができる。ここで、エレメントとは文書を構成する要素であって、たとえば文書、文および語が含まれる。

## 【0019】

受信部21は、この文書処理装置に外部からたとえば通信回線を介して送信される信号を受信する部分である。この受信部21は、たとえば電子文書である複数の文書を受信する。受信部21は、受信したデータを本体10に送る。

## 【0020】

出力部30は、この文書処理装置からの出力結果を表示するものである。この出力部30は、たとえば陰極線管(cathode ray tube;CRT)や液晶表示装置(liquid crystal display;LCD)から構成され、たとえば単数または複数のウィンドウを表示し、このウィンドウ上に文字、図形等を表示したりする。

## 【0021】

記録／再生部31は、この文書処理装置の制御部11の制御の下に、たとえばいわゆるフロッピーディスクのような記録媒体32に対して情報の記録／再生を行う。記録媒体32には、たとえば文書の語義に基づいて関連度を求め、この関連度に基づいて文書処理を実行するようなプログラムが記録されている。なお、この記録媒体32についてはさらに後述する。

## 【0022】

続いて、本実施の形態における文書について説明する。本実施の形態における文書は、ツリー状のタグ付けによる内部構造を有している。本実施の形態においては、図2に示すように、このタグ付けによる内部構造、文書、文、語彙エレメント等の各エレメント、通常リンク、参照・被参照リンク等が、タグとしてあらかじめ文書に付与されている。図中において、白丸“○”は文書の要素すなわちエレメントであり、最下位の白丸は文書における最小レベルの語に対応する、語彙エレメントである。また、実線は語、句、節、文等の文書の構造を示す通常リンク（normal link）である。破線は参照・被参照による係り受け関係を示す参照リンク（reference link）である。文書のタグ付けによる内部構造は、上位から下位への順序で、文書（document）、文書の下位であり段落の上位であるオプションのサブディビジョン（subdivision）、オプションの段落（paragraph）、文（sentence）、文の下位であるサブセンテンスセグメント（subsential segment）、・・・、最下位の語彙エレメントのような階層構造である。

## 【0023】

本実施の形態においては、文書のタグ付けによる内部構造として、多言語間に共通な意味的・語用論的タグを文書に付与することにより、文書の機械的な内容理解を可能にするようなタグ付けを採用している。タグとは、データに対してその属性を表すために付加される属性情報である。

## 【0024】

本実施の形態において文書のタグ付けによる内部構造は、HTML（Hyper Text Markup Language）と同様にXML（Extended Markup Language）の形式のタグである。すなわち、タグは、係り受け、たとえば代名詞の指示対象、多義語の意味のように統語（syntactic）・意味（semantic）等の情報を含んでいる。

## 【0025】

文章のタグ付けによる内部構造の一例を次に示すが、文章へのタグ付けはこの方法に限られないことはもちろんである。

## 【0026】

たとえば、“Time flies like an arrow.”という文については、

<文><名詞句 語義=“time0”>time</名詞句>  
 <動詞句><動詞 語義=“fly1”>flies</動詞>  
 <形容動詞句><形容動詞 語義=like0>like</形容動詞> <名詞句>an  
 <名詞 語義=“arrow0”>arrow</名詞></名詞句>  
 </形容動詞句></動詞句>.</文>

というようにタグ付けすることができる。ここで<文>、<名詞>、<名詞句>、<動詞>、<動詞句>、<形容動詞>、<形容動詞句>は、それぞれ一文、名詞、名詞句、動詞、動詞句、前置詞句、後置詞句を含む形容詞／形容詞句、形容詞句／形容動詞句のような文の統語構造（syntactic structure）を表している。タグは、エレメントの先端の前および終端の後に対応して配置される。ここでは、エレメントの終端の後ろに配置されるタグは、記号“/”により、文書の最小単位の要素、すなわちエレメントの終端であることを示している。このエレメントは統語的構成素、すなわち句、節、および文を示す。なお、語義=time0は、語timeの有する複数の意味、すなわち複数の語義のうちの第0番目の意味を指している。具体的には、timeには名詞と動詞があるが、ここではtimeが名詞であることを示している。たとえば、語“オレンジ”は色と果物の意味があるが、これらも語義によって区別することができる。

#### 【0027】

先に図2で説明したような文書のタグ付けによる内部構造は、図3のウィンドウ101に示すように、その統語構造を表示することができる。このウィンドウ101においては、右半分103が語彙エレメントを、左半分102が文の構造を示している。

#### 【0028】

このウィンドウ101には、タグ付けされた次に示すような文書が表示されている。この文書においても、タグによって統語構造が記述されている。次に示す文書は、「A氏のB会が終わったC市で、一部の大衆紙と一般紙がその写真報道を自主規制する方針を紙面で明らかにした。」についてのタグ付けによる内部構造を示すものである。

【0029】

<文書><文><形容動詞句 関係=“場所”><名詞句><形容動詞句 場所  
=“C市”>

<形容動詞句 関係=“主語”><名詞句 識別子=“B会”><形容動詞句  
関係 “位置”>A氏の</形容動詞句>B会</名詞句>が</形容動詞句>  
終わった</形容動詞句><地名 識別子=“C市”>C市</地名></名詞  
句>で、</形容動詞句><形容動詞句 関係=“主語”><名詞句 識別子=  
新聞”統語=“並列”><名詞句><形容動詞句>一部の</形容動詞句>大衆  
紙</名詞句>と<名詞>一般紙</名詞></名詞句>が</形容動詞句>  
<形容動詞句 関係=“目的語”><形容動詞句 関係=“内容” 主語=“新  
聞”><形容動詞句 関係=“目的語”><名詞句><形容動詞句><名詞 共  
参照=“B”>そ</名詞>の</形容動詞句>写真報道</名詞句>を</形  
容動詞句>

自主規制する</形容動詞句>方針を</形容動詞句>  
<形容動詞句 関係=“場所”>紙面で</形容動詞句>  
明らかにした。</文></文書>

【0030】

この文章においては、「一部の大衆紙と一般紙」のように、統語=“並列”  
は並列を表す。並列の定義は、係り受け関係を共有するということである。特に  
何も指定がない場合は、たとえば、<名詞句 関係=x><名詞>A</名詞>  
<名詞>B</名詞></名詞句> はAがBに依存関係のあることを表す。ま  
た、関係=xはこの<名詞句>エレメントの関係属性を表している。

【0031】

続いて、タグ付けにおける、統語、意味、修辭についての相互関係を記述する  
関係属性について説明する。

【0032】

主語、目的語、間接目的語のような文法機能、動作主、被動作者、受益者など  
のような主題役割、および理由、結果などのような修辭関係はこの関係属性によ  
って表示する。関係属性は関係=\*\*\* という形で表される。本実施の形態では、



比較的容易な文法機能、すなわち、主語、目的語、間接目的語のような文における当該語の機能について関係属性を記述する。

#### 【0033】

続いて、この文書処理装置の動作について、図4のフローチャートを参照して説明する。文書処理装置は、複数の文書について各文書の内容に関する特徴を表す特徴情報を含み、その文書の指標となるインデックスを作成する。そして、文書分類の分類モデルに基づいて、それぞれの文書のインデックスを参照することにより文書の自動的な分類を行う。分類モデルは、文書を分類する複数の分類項目から構成され、各分類項目は特徴を表す特徴情報を有している。

#### 【0034】

最初のステップS11においては、文書処理装置の受信部21は、外部から送信される複数の文書を受信する。文書処理装置は、受信部21にて受信された複数の文書を、制御部11の制御の下に、たとえばRAM14や記録／再生部31に記録する。文書は、図2に示したように、複数の要素すなわちエレメントからツリー状に構造化されたタグ付けによる内部構造を有している。

#### 【0035】

ステップS12においては、ユーザは、文書処理装置の表示部30に表示される文書を閲覧する。すなわち、文書処理装置の制御部11は、ユーザの希望に応じて、記憶する複数の文書から表示部30に文書を表示するように表示部30を制御する。文書処理装置の表示部30に表示する文書は、ユーザが入力部20に上記複数の文書のうちから所望の文書の選択を入力することにより任意に選択される。表示部30には、ユーザにより選択された文書の一部または全部の内容が、たとえばその領域の大きさを変更可能なウィンドウにより表示される。なお、ユーザが文書閲覧を行うこのステップS12は、ユーザの必要に応じて設けられる。また、図中においてこのステップS12が平行四辺形で表されているのは、ユーザが操作することに対応している。ステップS13においても同様である。

#### 【0036】

ここで、表示部30における表示の具体例について説明する。この具体例とは、ユーザが自由に文書を分類する分類項目であるカテゴリを設定、変更できるよ

うにしたものである。この具体例においては、ユーザが設定したカテゴリに応じて文書の自動分類が行われる。

#### 【0037】

このようなグラフィックユーザインターフェース (graphic user interface; GUI) の具体例は、図5に示すようになる。このGUI画面においては、操作ボタン302、“他のトピックス”を表示する第1の分類表示部303、“ビジネスニュース”を表示する第2の分類表示部304、“政治ニュース”を表示する第3の分類表示部305などを各分類項目が表示されている。“他のトピックス”は、“他のトピックス”、“ビジネスニュース”等の特定の分類項目に分類していない文書が分類される。各分類項目の表示部においては、文書のタイトルや文書の最初の部分が表示される。

#### 【0038】

また、このGUIにおいては、操作ボタン302は、画面のウィンドウの状態を初期の位置に戻すポジションリセット (position reset) と、文書の内容を閲覧するブラウザ (browser) を呼び出すブラウザのボタンと、このウィンドウから脱出するエグジット (exit) のボタンとを含んでいる。なお、上述の各分類表示部の大きさは固定的ではなく、所望の大きさに変更することができる。また、分類表示部のタイトルも自由に変更することができる。

#### 【0039】

この自動分類は、ユーザの個別の要求に応じて分類項目を決めることにより、ユーザの関心に応えたり、ユーザが文書を探すときの効率の向上を図るものである。

#### 【0040】

ステップS13においては、ユーザは、ステップS12において文書処理装置の表示部30にて閲覧した複数の文書について、この複数の文書を分類する分類項目、いわゆるカテゴリを作成し、この分類項目に上記複数の文書を分類する。文書処理装置においては、文書を分類する分類項目の設定は、たとえば分類項目の数に対応して分割された領域を有するウィンドウについて、所望の分類項目を追加したり、あるいは変更や削除をすることにより行われる。複数の文書の分類項

目への分類は、たとえば文書の一部や分類項目のタイトルが表示され、このタイトルに対応する領域が設けられたウインドウにおいて、たとえば画面上に表示されたアイコンをカーソルに連動するマウスをクリックしてドラッグすることにより、文書を所望の領域に移動して各文書を分類する。このような、文書の分類作成および分類操作は、ユーザが表示部 30 の表示を参照しながら入力部 20 に入力することにより行われる。作成された分類項目および分類操作の結果は、制御部 11 の制御の下に、たとえば RAM 14 に記録される。なお、文書の分類項目の作成および文書の分類の操作の詳細については、さらに後述する。

#### 【0041】

ステップ S 14 においては、文書処理装置の制御部 11 は、ステップ S 13 において行われた分類項目の作成と、この分類項目に応じた分類操作に基づいて、分類モデルの作成を実行する。文書処理装置は、たとえば RAM 14 に記録されたステップ S 13 における分類項目および分類操作の結果を読み出す。そして、文書処理装置の制御部 11 は、この結果に基づいて、各分類項目に分類された上記複数の文書について、各分類項目に特徴的な固有名詞、固有名詞以外の語義、分類された文書のアドレスを集めて、分類モデルを生成する。ここで、固有名詞以外の場合に語そのものではなく語義を用いるのは、同じ語でも複数の意味を有することがあるからである。そして、文書処理装置の制御部 11 は、このように作成した分類モデルをたとえば RAM 14 に記憶する。なお、分類モデルの作成の詳細については、さらに後述する。

#### 【0042】

以上の一連の行程により、文書を分類する基準となる分類モデルが作成された。文書処理装置は、この分類モデルを基準として、文書を自動的に分類することができる。文書処理装置が行う、新たに受信した文書の自動的な分類の動作について、図 6 を参照して説明する。

#### 【0043】

文書処理装置において、外部からたとえば通信回線を介して受信部 21 に新たな文書が送信されると、文書処理装置はこの文書を受信する。文書処理装置における文書の受信の動作については、上述したステップ S 11 で詳しく述べたので

、ここでは説明を省略する。受信した文書は、たとえばRAM 14や記録／再生部31に記録される。

【0044】

ステップS22においては、文書処理装置の制御部11は、たとえばRAM 14や記録／再生部31に記録されたステップS21で受信した文書を読み出す。制御部11は、この新たな文書から各文書の特徴を表す語を抽出することによりその文書の指標、すなわちインデックスの作成を行う。そして、制御部11は、各文書についてのインデックスをたとえばRAM 14に記録する。なお、このインデックス作成の詳細については、さらに後述する。

【0045】

ステップS23においては、文書処理装置の制御部11は、分類モデルに基づいて、インデックスを附された各文書を上述のステップS13において作成した複数の分類項目の一つに分類する。そして、制御部11は、分類の結果をたとえばRAM 14に記録する。なお、このような文書の自動分類の詳細については、さらに後述する。

【0046】

ステップS24においては、文書処理装置の制御部11は、たとえばRAM 14に記録されたステップS23での新たな文書の自動分類の結果に基づいて、分類モデルを更新する。制御部11は、更新した分類モデルをたとえばRAM 14に記録する。

【0047】

上述したタグ付けによる内部構造を有する文書は、文書処理装置の受信部21に外部から送信される。この文書は、たとえばデジタル符号化されたいわゆる電子文書である。文書処理装置は、このような文書をたとえばRAM 14や記録／再生部31に記録する。ユーザは、文書処理装置の記録する複数の文書から、任意の文書を表示部30に表示して閲読することができる。

【0048】

表示部30における文書の表示は、たとえば大きさを変更することができるウィンドウ上に表示することができる。また、文書の表示と共に、または文書の表

示に代えて要約を表示することができる。さらに、複数の文書をウィンドウにより並べて表示したり、複数のウィンドウを重ねて表示することができる。

【0049】

文書処理装置の制御部 11 は、このように表示部 30 に表示された文書について、ユーザの入力にしたがい、各種の処理を実行する。ユーザによる、文書についての入力、表示部 30 に表示されるカーソルに連動する入力部 20 のマウスをクリックすることにより表示部 30 における所定の領域を指定したり、入力部 20 のキーボードによりキーワードを入力したりすることにより行う。

【0050】

次に、図 4 および図 6 の処理の詳細について説明する。

【0051】

最初に、S 22 の文書の特徴を発見してインデックスを作る手順について詳細に説明する。このインデックスとは、文書の特徴を表す語を各文書について抽出して指標としたものである。文書の特徴を発見してインデックスを作成する手順は、文書処理装置の制御部 11 の制御の下に、図 7 のフローチャートに示す一連の手順により実行される。すなわち、以下の手順は語義の関連度を算出して、この関連度に基づいてインデックスの作成を行うものである。

【0052】

最初のステップ S 31 においては、制御部 11 は、図 6 のステップ S 21 において受信した文書についてその文書内で活性拡散を実行し、文書内の各要素の活性値を拡散する。制御部 11 による文書の各要素への活性値の拡散処理は、後に詳細に説明する。制御部 11 は、活性拡散の結果として得られた活性値をたとえば RAM 14 に記録する。

【0053】

ステップ S 32 においては、制御部 11 は、ステップ S 11 で得られた活性値に基づいて、活性値が予め設定された閾値を超える要素を抽出する。制御部 11 は、このように抽出した要素をたとえば RAM 14 に記録する。

【0054】

ステップ S 33 においては、制御部 11 は、たとえば RAM 14 からステップ

S 3 2 にて抽出したエレメントを読み出す。そして、制御部 1 1 は、このエレメントからすべての固有名詞を取り出してインデックスに加える。固有名詞は語義を持たず、辞書に載っていないなどの特殊の性質を有するので、固有名詞以外の語とは別に扱う。固有名詞であるか否かは、たとえば文書に付加されたタグにより識別される。たとえば、図 3 に示したタグ付けによる内部構造においては、“A 氏”、“B 会”および“C 市”は固有名詞である。そして、制御部 1 1 は、取り出した固有名詞をインデックスに加え、その結果をたとえば R A M 1 4 に記録する。

#### 【0055】

ステップ S 3 4 においては、制御部 1 1 は、たとえば R A M 1 4 からステップ S 3 2 にて抽出したエレメントから固有名詞以外の語義を取り出してインデックスに加え、その結果を R A M 1 4 に記録する。ここでの語義とは、語の有する複数の意味の内の選択された一つである。本実施の形態においては、語義についてもタグ付けにより記述されている。

#### 【0056】

このように、文書の特徴を発見してインデックスを作成する手順は、複数のエレメントから構成されるタグ付けによる内部構造を有する文書の特徴を発見して、その特徴を配列したインデックスを作るものである。すなわち、タグ付けによる内部構造に基づいて上記文書について活性拡散をすることにより、各語彙エレメントの活性値を拡散し、拡散後の活性値が所定の閾値より大きい語彙エレメントを抽出する。そして、その語彙エレメントについて、固有名詞または語義をインデックスに追加する。

#### 【0057】

なお、上述のインデックスには、文書の特徴を表す語と共に、その文書のアドレスを含めることもできる。すなわち、インデックスはその文書を代表するような特徴を表す語を含むので、所望の文書を参照する際の指標とすることができる。このインデックスは、上述したように文書の自動的な分類に利用することができるが、これについての詳細はさらに後述する。

【0058】

ここで、インデックスの具体例を示す。

【0059】

<インデックス 日付=“AAAA/BB/CC” 時刻=“DD:EE:FF” 文書アドレス=“1234”>

<要約>減税規模、触れず-X首相の会見</要約>

<語 語義=“0003” 活性値=“140.6”>触れず</語>

<語 語義=“0105” 識別子=“X” 活性値=“140.6”>首相</語>

<名前 識別子=“X” 語 語義=“6103” 活性値=“140.6”>X首相</語>

<語 語義=“5301” 活性値=“140.6”>求めた</語>

<語 語義=“2350” 識別子=“X” 活性値=“140.6”>首相</語>

<語 語義=“9582” 活性値=“140.6”>強調した</語>

<語 語義=“2595” 活性値=“140.6”>触れる</語>

<語 語義=“9472” 活性値=“140.6”>予告した</語>

<語 語義=“4934” 活性値=“140.6”>触れなかった</語>

<語 語義=“0178” 活性値=“140.6”>釈明した</語>

<語 語義=“7248” 識別子=“X” 活性値=“140.6”>私</語>

<語 語義=“3684” 識別子=“X” 活性値=“140.6”>首相</語>

<語 語義=“1824” 活性値=“140.6”>訴えた</語>

<語 語義=“7289” 活性値=“140.6”>見せた</語>

</インデックス>

【0060】

このインデックスにおいては、<インデックス>および</インデックス>は、インデックスの始端および終端を、<日付>および<時刻>はこのインデックスが作成された日付および時刻を、<要約>および</要約>はこのインデックスの内容の要約の始端および終端を示している。また、<語>および</語>は語の始端および終端を、それぞれ示している。語義=“0003”は、その語義が、複数の語義のうちの第3番目であることを示している。他にも同様である。

## 【0061】

続いて、タグ付けによる内部構造に基づいて、ステップS31で行う活性拡散によりエレメントの活性値を拡散する方法について説明する。なお、この活性拡散は、後述する図11におけるステップS62においても実行される。

## 【0062】

タグ付けによる内部構造を与えられた文書においては、活性拡散と呼ばれる処理を行うことにより、各エレメントにタグ付けによる内部構造に応じた活性値を付与することができる。活性拡散は、活性値の高いエレメントと関わりのあるエレメントにも高い活性値を与えるような処理である。この活性値は、タグ付けによる内部構造に応じて決定されるので、タグ付けによる内部構造を考慮した文書の分析に利用することができる。

## 【0063】

活性拡散は、図8のフローチャートに示す一連の行程にしたがって、文書処理装置の制御部11の制御の下に実行される。

## 【0064】

最初のステップS41においては、文書内のエレメントの活性値の初期化が行われる。すなわち、制御部11は、語彙エレメントを除いたすべてのエレメントと語彙エレメントに対して活性値の初期値を割り当てる。たとえば、活性値の初期値として語彙エレメントを除いたすべてのエレメントに1を、語彙エレメントに零を、それぞれ割り当てればよい。また、制御部11は、各エレメントの活性値の初期値に均一ではない値を割り当てることにより、活性拡散の結果得られた活性値の初期値の偏りを反映することができる。たとえば、ユーザが関心を有するエレメントに対しては、活性値の初期値を高く設定することにより、ユーザの関心を反映した活性値を得ることができる。

## 【0065】

参照・被参照関係のエレメントを連結する参照・被参照リンクとそれ以外の通常リンクに関しては、エレメントを連結するリンクの端点の活性値である端点活性値を0に設定する。制御部11は、このように付与した活性値の初期値をたとえばRAM14に記録する。



## 【0066】

エレメントとエレメントの連結は、たとえば図9に示すようになる。この図においては、文書を構成するエレメントとリンクの構造の一部として、エレメント $E_i$ およびエレメント $E_j$ が示されている。エレメント $E_i$ とエレメント $E_j$ とは、活性値 $e_i$ および $e_j$ をそれぞれ有し、リンク $L_{ij}$ にて接続されている。リンク $L_{ij}$ のエレメント $E_i$ に接続する端点は $T_{ij}$ 、エレメント $E_j$ に接続する端点は $T_{ji}$ である。エレメント $E_i$ は、リンク $L_{ij}$ により接続されるエレメント $E_j$ の他に、リンク $L_{ik}$ 、 $L_{il}$ および $L_{im}$ によって図示しないエレメント $E_k$ 、 $E_l$ および $E_m$ にそれぞれ接続している。エレメント $E_j$ は、リンク $L_{ji}$ により接続されるエレメント $E_i$ の他に、リンク $L_{jp}$ 、 $L_{jq}$ および $L_{jr}$ によって図示しないエレメント $E_p$ 、 $E_q$ および $E_r$ にそれぞれ接続している。

## 【0067】

ステップS42においては、文書処理装置の制御部11は、文書を構成するエレメント $E_i$ を計数するカウンタの初期化を行う。すなわち、エレメントを計数するカウンタのカウント値 $i$ を1に設定する。すなわち、このカウンタは、第1番目のエレメント $E_1$ を参照している。

## 【0068】

ステップS43においては、文書処理装置の制御部11は、カウンタが参照するエレメントについて、活性値を計算するリンク処理を実行する。このリンク処理については、さらに後述する。

## 【0069】

ステップS44においては、文書処理装置の制御部11は、文書中のすべてのエレメントについて活性値の計算が完了したか否かを判断する。そして、制御部11は、文書中のすべてのエレメントについて活性値の計算が完了したときには“YES”としてステップS45に処理を進め、文書中のすべてのエレメントについて活性値の計算が完了していないときには“NO”としてステップS47に処理を進める。

## 【0070】

具体的には、制御部11は、カウンタにて計数されている活性値の計算がなさ

れたエレメントを参照するカウンタのカウント値  $i$  が、文書の含むエレメントの総数に達したか否かを判断する。そして、制御部 11 は、カウンタのカウント値  $i$  が文書に含まれるエレメントの総数に達したときには、すべてのエレメントが計算済みとしてステップ S45 に処理を進め、カウンタのカウント値  $i$  が文書に含まれるエレメントの総数に達していないときにはすべてのエレメントについて計算が終了していないとしてステップ S47 に処理を進める。

## 【0071】

ステップ S47 においては、文書処理装置の制御部 11 は、カウンタのカウント値  $i$  を 1 増加させて、カウンタのカウント値を  $i+1$  とする。このことにより、カウンタは  $i+1$  番目のエレメント、すなわち次のエレメントを参照する。そして、処理はステップ S43 に戻り、端点活性値の計算およびこれに続く一連の行程が、次の  $i+1$  番目のエレメントについて実行される。

## 【0072】

具体的には、制御部 11 は、エレメントを計数するカウンタのカウント値  $i$  を 1 増加する。このことにより、カウンタはステップ S43 で活性値が計算されたエレメントの次のエレメントを参照することになる。

## 【0073】

ステップ S45 においては、文書処理装置の制御部 11 は、文書に含まれるすべてのエレメントの活性値の変化分、すなわち新たに計算された活性値の元の活性値に対する変化分について、文書に含まれるすべてのエレメントについて平均値を計算する。

## 【0074】

文書処理装置の制御部 11 は、たとえば RAM 14 に記録された元の活性値と新たに計算した活性値を、文書に含まれるすべてのエレメントについて読み出す。制御部 11 は、新たに計算した活性値の元の活性値に対するそれぞれの変化分の総和を文書に含まれるエレメントの総数で除することにより、すべてのエレメントの活性値の変化分の平均値を計算する。制御部 11 は、このように計算したすべてのエレメントの活性値の変化分の平均値を、たとえば RAM 14 に記録する。

## 【0075】

ステップS46においては、制御部11は、ステップS49で計算したすべてのエレメントの活性値の変化分の平均値が、予め設定された閾値以内であるか否かを判断する。そして、制御部11は、上記変化分が閾値以内であると“YES”としてこの一連の行程を終了する。上記制御部11は、上記変化分が閾値以内でないときには“NO”として、ステップS42にてカウンタのカウント値iを1に設定して文書のエレメントの活性値を計算する一連の行程を再び実行する。この一連の行程にて構成されるステップS42からステップS44に至るループが繰り返される毎に上記変化分は徐々に減少する。

## 【0076】

続いて、ステップS43にて実行される活性値を計算するリンク処理について、図10に示すフローチャートを参照して説明する。

## 【0077】

ステップS51においては、文書処理装置の制御部11は、図9に示すように、文書を構成するエレメント $E_j$ を計数するカウンタの初期化を行う。すなわち、エレメントを計数するカウンタのカウント値jを1に設定する。すなわち、このカウンタは、第1番目のエレメント $E_j$ を参照している。

## 【0078】

ステップS52においては、エレメント $E_i$ と $E_j$ を接続するリンク $L_{ij}$ においては、制御部11は、タグを参照することにより、そのリンク $L_{ij}$ が通常リンクであるか否かを判断する。制御部11は、リンク $L_{ij}$ について、そのリンクが、語に対応する語彙エレメント、文に対応する文エレメント、段落に対応する段落エレメントなどの間の関係を示す通常リンクと、参照・被参照による係り受けの関係を示す参照リンクのいずれであるかを判断する。これは、図3の“関係”を参照することで判断することができる。制御部11は、そのリンクが通常リンクのときには“YES”としてステップS53に処理を進め、そのリンクが参照リンクのときには“NO”としてステップS54に処理を進める。

## 【0079】

ステップS53においては、通常リンク $L_{ij}$ に対して、そのリンクの端点の活

性値を計算する処理が行われる。この端点活性値の計算について、図9を参照して説明する。

#### 【0080】

ここでは、ステップS52における判別により、リンク $L_{ij}$ は通常リンクであることが明らかになっている。通常リンク $L_{ij}$ に関して、エレメント $E_i$ に接続する端点 $T_{ij}$ の端点活性値 $t_{ij}$ は、このリンク $L_{ij}$ を除いたエレメント $E_i$ に接続するすべてのリンク $L_{ik}$ 、 $L_{il}$ および $L_{im}$ の端点活性値 $t_{ik}$ 、 $t_{il}$ および $t_{im}$ と、このエレメント $E_i$ がリンク $L_{ij}$ により接続するエレメント $E_j$ の活性値 $e_j$ を加算し、この加算で得た値を文書に含まれるエレメントの総数で除することにより求められる。

#### 【0081】

エレメント $E_{ij}$ の端点 $T_{ij}$ の端点活性値は、端点 $T_{ij}$ を一端とするリンク $L_{ij}$ が通常リンクの場合、リンク $L_{ij}$ の他端が接続されているエレメント $E_j$ の端点の端点活性値のうちそのリンク $L_{ij}$ と接続されている端点 $T_{ji}$ を除いたすべての端点の端点活性値、およびそのリンク $L_{ij}$ が接続されるエレメント $E_j$ の活性値 $e_j$ の和を文書全体に含まれるエレメントの総数で除することにより得られる。このような手順により、活性拡散における活性値の収束が保証されることになる。

#### 【0082】

文書処理装置の制御部11は、たとえばRAM14に記録されたデータから、必要な端点活性値および活性値を読み出す。制御部11は、読み出された端点活性値および活性値について、上述のようにその通常リンクと接続された端点の端点活性値を計算する。そして制御部11は、このように計算した端点活性値を、たとえばRAM14に記録する。

#### 【0083】

ステップS54においては、参照リンクに対して、そのリンクの端点の活性値を計算する処理が行われる。

#### 【0084】

ステップS52における判別により、リンク $L_{ij}$ は参照リンクであることが明

らかになっている。通常リンク  $L_{ij}$  に関して、エレメント  $E_i$  に接続する端点  $T_{ij}$  の端点活性値  $t_{ij}$  は、このリンク  $L_{ij}$  を除いたエレメント  $E_i$  に接続するすべてのリンク  $L_{ik}$ 、 $L_{il}$  および  $L_{im}$  の端点活性値  $t_{ik}$ 、 $t_{il}$  および  $t_{im}$  と、このエレメント  $E_i$  がリンク  $L_{ij}$  により接続するエレメント  $E_j$  の活性値  $e_j$  を加算することにより求められる。

## 【0085】

エレメント  $E_i$  の端点  $T_{ij}$  の端点活性値は、端点  $T_{ij}$  を一端とするリンク  $L_{ij}$  が参照リンクの場合、リンク  $L_{ij}$  の他端が接続されているエレメント  $E_j$  の端点の端点活性値のうちそのリンク  $L_{ij}$  と接続されている端点  $T_{ji}$  を除いたすべての端点の端点活性値、およびそのリンク  $L_{ij}$  が接続されるエレメント  $E_j$  の活性値  $e_j$  の和を取ることににより得られる。

## 【0086】

文書処理装置の制御部 11 は、たとえば RAM 14 に記録されたデータから、必要な端点活性値および活性値を読み出す。制御部 11、読み出された端点活性値および活性値について、上述のようにその参照リンクの端点活性値を計算する。そして制御部 11 は、このように計算した端点活性値を、たとえば RAM 14 に記録する。

## 【0087】

ステップ S53 における通常リンクの処理、およびステップ S54 における参照リンクの処理は、ステップ S42 のカウンタのカウンタ値  $i$  により参照されているエレメント  $E_i$  に接続するすべてのエレメント  $E_j$  についてのリンク  $L_{ij}$  に対して実行される。

## 【0088】

ステップ S55 においては、文書処理装置の制御部 11 は、ステップ S53 またはステップ S54 での計算に基づいて、エレメント  $E_i$  の端点活性値を計算する。制御部 11 は、この計算により得られた端点活性値をたとえば RAM 14 に記録する。

## 【0089】

ステップ S56 においては、エレメント  $E_i$  に接続するすべてのリンクについ

て端点活性値  $t_{ij}$  が計算されたか否かが判別される。そして、すべてのリンクについて端点活性値が計算されているときには“YES”としてステップ S57 に進み、すべてのリンクについて端点活性値が計算されていないときには“NO”としてステップ S58 に進む。

#### 【0090】

ステップ S57 においては、S56 にてエレメント  $E_i$  のすべてのリンク  $L_{ij}$  について端点活性値  $t_{ij}$  が求められたことが判別されたので、エレメント  $E_i$  の活性値  $e_i$  の更新を実行する。

#### 【0091】

エレメント  $E_i$  の活性値  $e_i$  の新たな値すなわち更新値は、エレメント  $E_i$  のすべての端点の活性値の和  $e_i' = e_i + \sum t_{ij}'$  を取ることにより求められる。ここで、“’”は、新たな値という意味である。このように、活性値は、そのエレメントに接続するすべてのリンクについて、そのエレメントに接続する端点の端点活性値の総和となる。

#### 【0092】

文書処理装置の制御部 11 は、たとえば RAM14 に記録されたデータから必要な端点活性値  $t_{ij}$  を読み出す。制御部 11 は、上述したような計算を実行し、そのエレメント  $E_i$  の活性値  $e_i$  を算出する。そして、制御部 11 は、計算した新たな活性値  $e_i$  をたとえば RAM14 に記録する。

#### 【0093】

次に、上述した活性値に基づいて行う語義の関連度の計算について、図 11 に示すフローチャートを参照して説明する。語義の関連度の計算は、図 4 および図 6 に示す処理を行う前にあらかじめ行う前処理であるから、一度実行すればよい。

#### 【0094】

最初のステップ S61 において、制御部 11 は、電子辞書内の語の語義の説明を用い、辞書を使って語義のネットワークを作成する。すなわち、辞書における各語義の説明と、この説明中に現れる語義との参照関係から、上述したような語義のタグ付けによる構造のネットワークを作成する。これは、最上位のエレメン

トを辞書として、図2に示したようなタグ付けによる内部構造を構成することに相当する。制御部11は、RAM14に記録した語義とその説明を順に読み出して、ネットワークを作成する。制御部14は、このようにして作成した語義のネットワークをたとえばRAM14や記録／再生部31に記録する。

【0095】

なお、この辞書は、たとえば通信回線から受信部21にて受信することができる。また、たとえばCD-ROMなどの記録媒体32によって提供され、記録／再生部31で再生することができる。

【0096】

ステップS62において、ステップS61で作成された語義のネットワーク上で、上述した活性拡散を行う。この活性拡散により、各語義の活性値は、上記辞書により与えられたタグ付けによる内部構造に応じて更新される。

【0097】

ステップS63においては、ステップS61で作成された語義のネットワークを構成する一つの語義 $s_i$ を選択し、ステップS64においては、この語義 $s_i$ に対応する語彙エレメント $E_i$ の活性値 $e_i$ の初期値を適当の変化させ、このときの活性値の差分 $\Delta e_i$ を計算する。

【0098】

ステップS65においては、ステップS64におけるエレメント $E_i$ の活性値 $e_i$ の初期値の変化に対応する、語義 $s_j$ に対応するエレメント $E_j$ の活性値 $e_j$ の差分 $\Delta e_j$ を求める。ステップS66においては、ステップS65で求めた差分 $\Delta e_j$ をステップS64で求めた $\Delta e_i$ で除した商 $\Delta e_j / \Delta e_i$ を、語義 $s_j$ の語義 $s_j$ に対する関連度とする。ある語義の活性値をステップS64で変えたのに応じて、関連する語の活性値が変わることとなる。

【0099】

ステップS67においては、語義 $s_i$ と $s_j$ とのすべての組について関連度の演算が終了したか否かについて判断する。そして、すべての語義の組について関連度の演算が終了したときには“YES”として、この一連の処理を終了する。すべての語義の組について関連度の演算が終了していないときには、“NO”とし

て、ステップ S 6 3 にもどり、関連度の演算が終了していない組について関連度の演算を継続する。

#### 【0100】

このように計算された関連度は、図 1 2 に示すように、それぞれの語義と語義の間に定義される。この語義の表においては、関連度は正規化され、0 から 1 までの値をとる。すなわち、この語義の表においては“コンピュータ”、“テレビ”、“VTR”の間の相互の関連度が示されている。“コンピュータ”と“テレビ”の関連度は 0.55、“コンピュータ”と“VTR”の関連度は 0.25、“テレビ”と“VTR”の関連度は 0.60 である。制御部 11 は、このように作成した関連度をたとえば RAM 14 に記憶する。

#### 【0101】

ステップ S 6 3 からステップ S 6 7 のループにおいては、制御部は、必要な値をたとえば RAM 14 や記録／再生部 31 から順に読み出して、上述したように関連度を計算する。制御部 11 は、計算した関連度をたとえば RAM 14 や記録／再生部 31 に記録する。

#### 【0102】

次に、上述したように算出された関連度を用いた文書分類について説明する。この関連度を利用した文書分類は、先に説明した図 5 の GUI における文書分類に用いられる。

#### 【0103】

関連度による文書分類は、各分類項目の特徴を示す分類モデルを参照して関連度に基づいて行われる。分類モデルとは、各分類項目に特徴的な、固有名詞、固有名詞以外の語義、アドレスなどを含んで構成される。たとえば、図 1 3 に示す分類モデルは、各分類項目、いわゆるカテゴリに対して、固有名詞、固有名詞以外の語義、アドレスの欄を有している。この分類モデルにおいては、分類項目は“スポーツ”、“社会”、“コンピュータ”、“植物”、“美術”および“美術”の各項目から構成されている。これらの分類項目に対応する固有名詞として、“A 氏”、“B 社”、“C 社”および“G 社”、“D 種”、“E 氏”、“F 氏”がそれぞれ示されている。上記分類項目に対応する固有名詞以外の語義として、



“野球”および“グラウンド”、“労働”および“雇用”、“モバイル”、“桜1”および“オレンジ1”、“桜2”および“オレンジ2”、“桜3”がそれぞれ示されている。また、上記分類項目に対応するアドレスとして、“12345”、“22222”、“33333”、“44444”、“55555”、“66666”がそれぞれ示されている。なお、“桜1”、“桜2”および“桜3”は“桜”の第1の語義(11111)、第2の語義(11112)および第3の語義(11113)を示している。また、“オレンジ1”および“オレンジ2”は、“オレンジ”の第1および第2の語義を示している。

#### 【0104】

各分類項目の分類モデルは、タグ付けによる内部構造による活性値に基づいて抽出される。上述したように、文書処理装置の制御部11は、ステップS32において活性値が所定の閾値を超えるエレメントを抽出し、ステップS33においてこのエレメントからすべての固有名詞を取り出してインデックスに加え、ステップS34においては固有名詞以外の語義を取り出してインデックスに加える。このように分類モデルの特徴の欄は、たとえば上述の手順により生成されたインデックスを分類項目ごとにまとめたものである。

#### 【0105】

図6におけるステップS23で行われる文書の自動分類は、このような分類モデルを参照して、図14のフローチャートに示す一連の手順に従って、語義の関連度に基づいて行われる。

#### 【0106】

ステップS71においては、制御部11は、分類モデルの各分類項目 $C_i$ に含まれる固有名詞の集合と、ステップS62において文書から抽出されインデックスに入れられた語のうちの固有名詞の集合とについて、これらの共通集合の数を $P(C_i)$ とする。そして、制御部11は、このようにして算出した数 $P(C_i)$ をたとえばRAM14に記録する。

#### 【0107】

ステップS72においては、制御部11は、その文書のインデックス中の語義と各分類項目 $C_i$ に含まれる語義との関連度を図12の語義の表を参照し、語義

の関連度の総和  $R(C_i)$  を演算する。すなわち、制御部 11 は、分類モデルにおける固有名詞以外の語について、ステップ S61 で算出した関連度の総和  $R(C_i)$  をとる。そして、制御部 11 は、算出した関連度の総和  $R(C_i)$  をたとえば RAM14 に記録する。

## 【0108】

ステップ S73 においては、項目  $C_i$  に対する文書の関連度を、

$$Rel(C_i) = mP(C_i) + nR(C_i)$$

と定義する。ここで、係数  $m$ 、 $n$  は定数で、それぞれの値の関連度への貢献の度合いを表すパラメータである。制御部 11 は、ステップ S33 で算出した共通集合の数  $P(C_i)$  およびステップ S64 で算出した語義の関連度の総和  $R(C_i)$  をたとえば RAM14 から読み出し、上述の式に当てはめて文書の関連度  $Rel(C_i)$  を算出する。なお、これらの係数  $m$ 、 $n$  の値としては、たとえば  $m=10$ 、 $n=1$  とすることができる。そして、制御部 11 は、このように算出した文書の関連度  $Rel(C_i)$  をたとえば RAM14 に記録する。

## 【0109】

係数  $m$  および  $n$  の値は、統計的手法を使って推定することもできる。すなわち、制御部 11 は、複数の係数  $m$  および  $n$  の組について文書の関連度  $Rel(C_i)$  が与えられると、上記係数を最適化により求めることができる。

## 【0110】

ステップ S74 においては、制御部 11 は、項目  $C_i$  に対する関連度  $Rel(C_i)$  が全項目中最大で、その関連度の値がある閾値を超えているとき、分類項目  $C_i$  に文書を分類する。すなわち、制御部は、複数の項目についてそれぞれ文書の関連度  $Rel(C_i)$  を作成し、最大の関連度  $Rel(C_i)$  が閾値を超えているときには、文書を上記項目に分類  $C_i$  する。最大の関連度  $Rel(C_i)$  が閾値を超えていないときには、文書の分類は行わない。

## 【0111】

このように、文書中に含まれる語義間の関連度の計算とそれに基づく文書の分類の手順は、複数のエレメントから構成されるタグ付けによる内部構造を有する文書进行处理し、この文書を複数の分類項目の内一つの分類項目に分類する。す

なわち、この手順は、文書と各分類項目との関連度を算出し、算出された関連度に基づいて上記文書を分類する分類項目を決定する。

#### 【0112】

ここで、文書を分類する分類項目は、文書から抽出された固有名詞および／または語義を含む分類モデルによって特徴づけられる。そして、このような分類モデルを用いて、各分類項目の分類モデルに含まれる固有名詞および文書から抽出された固有名詞についての共通の数を算出し、各分類項目の上記分類モデルに含まれる語義に対する上記文書の関連度の総和を算出する。さらに、文書に含まれる固有名詞と関連度に基づいて抽出された固有名詞において重複する固有名詞の数と、語義の関連度の総和とに基づいて上記文書を分類する分類項目を決定する。この語義の関連度は、上述したようなタグ付けによる内部構造に基づいて決定される。

#### 【0113】

文書の分類は、たとえば共通する固有名詞の数および語義の関連度の線形結合が最大となって、所定の閾値を越えるような項目に対して行われる。このような共通する固有名詞の数および語義の関連度の線形結合の係数は、文書と分類項目の関連の大きさから、上述したように統計的に決定することができる。

#### 【0114】

次に、文書処理装置の記録／再生部 31 において記録／再生される記録媒体 32 について説明する。記録媒体には、複数のエレメントからタグ付けによる内部構造を有する文書进行处理する文書処理プログラムが記録されている。この記録媒体 32 としては、情報の記録／再生が可能なたとえばフロッピーディスクが利用される。

#### 【0115】

記録媒体 32 において、文書処理プログラムは、エレメントの最小単位である語義ごとに他の語義を参照する辞書を用い、語義の参照関係を組織する参照関係組織処理と、参照関係組織処理で組織された参照関係の組織の構造に基づいて語義にそれぞれ活性値を付与する活性値付与処理と、活性値付与処理で上記語義に付与された活性値に、参照関係の組織の構造に基づいた演算を施すことにより、

語義に新たに活性値を付与する活性値付与処理と、活性値付与処理で一の語義に付与された活性値を独立変数とし、活性値演算処理で他の語義に付与された活性値を従属変数とし、活性演算処理で他の語義に付与された活性値の微分を活性値付与工程で一の語義に付与された活性値の微分で除した微分商を一の語義と他の語義の関連度として演算する関連度演算処理との各処理工程を有するものである。

#### 【0116】

記録媒体 32 においては、参照関係組織処理で用いられる辞書の各語義にはその語義の属性を示す属性情報が付与され、上記参照関係の組織は上記属性情報に基づいて作成される。

#### 【0117】

記録媒体 32 において、文書処理プログラムは、エレメントの最小単位である語義の間の相互の関連度を算出する関連度算出処理と、文書を分類する複数の分類項目について、分類項目の特徴をあらわす語義を含んでなる分類モデルを用い、各分類モデルの含む語義との関連度に基づいて文書を分類する文書分類処理との各処理工程を有する。

#### 【0118】

記録媒体 32 において、関連度算出処理は、エレメントの最小単位である語義ごとに他の語義を参照する辞書を用い、上記語義の参照関係を組織する参照関係組織処理と、参照関係組織処理で組織された参照関係の組織の構造に基づいて語義の構造に基づいた演算を施すことにより、語義に新たに活性値を付与する活性値付与処理と、活性値付与処理で一の語義に付与された活性値を独立変数とし、活性値演算処理で他の語義に付与された活性値を従属変数とし、活性値演算処理で他の語義に付与された活性値の微分を活性値付与処理で一の語義に付与された活性値の微分で除した微分商を一の語義と他の語義の関連度として演算する関連度演算処理との各処理工程を有する。

#### 【0119】

なお、本実施の形態においては、文書へのタグ付けの方法の一例を示したが、本発明がこのタグ付けの方法に限定されないことはもちろんである。また、本実

施の形態においては、文書処理装置の受信部 21 に外部から文書が送信されたとしたが、本発明はこれに限定されない。たとえば、上記文書は、文書処理装置の ROM 13 に書き込まれていたり、記録／再生部 31 において記録媒体 32 から読み出されてもよい。

#### 【0120】

また、上述の実施の形態においては、文書処理装置の表示部 30 に表示された文書から所望の要素を選択するデバイスとしてマウスを例示したが、本発明がこれに限定されないことはいうまでもない。文書処理装置における要素の入力には、タブレット、ライトペン等の他のデバイスを利用することができる。

#### 【0121】

##### 【発明の効果】

上述のように、本発明に係る文書処理方法および装置ならびに記録媒体を利用することにより、語義の関連度を算出することができる。この語義の関連度を利用することにより、語義の関連度に基づいて、ユーザの興味を反映した文書の自動分類のような文書処理を実行することができるようになる。語義の関連度に基づく文書処理は、自動的に実行することができるので、文書処理の際のユーザの負担を軽減する。

##### 【図面の簡単な説明】

##### 【図 1】

本実施の形態を適用した文書処理装置の構成を示すブロック図である。

##### 【図 2】

文書のタグ付けによる内部構成を示す図である。

##### 【図 3】

文書のタグ付けによる内部構成を表示したウィンドウを示す図である。

##### 【図 4】

本実施の形態を適用した文書処理装置の動作を示すフローチャートである。

##### 【図 5】

文書の自動分類を行う GUI を示す図である。

【図 6】

文書を自動分類するフローチャートである。

【図 7】

文書の特徴を発見してインデックスを作成するフローチャートである。

【図 8】

活性拡散を示すフローチャートである。

【図 9】

活性拡散の処理を説明する図である。

【図 10】

活性拡散のリンク処理のフローチャートである。

【図 11】

語義の関連度の計算のフローチャートである。

【図 12】

語義の関連度の表を示す図である。

【図 13】

分類モデルの表を示す図である。

【図 14】

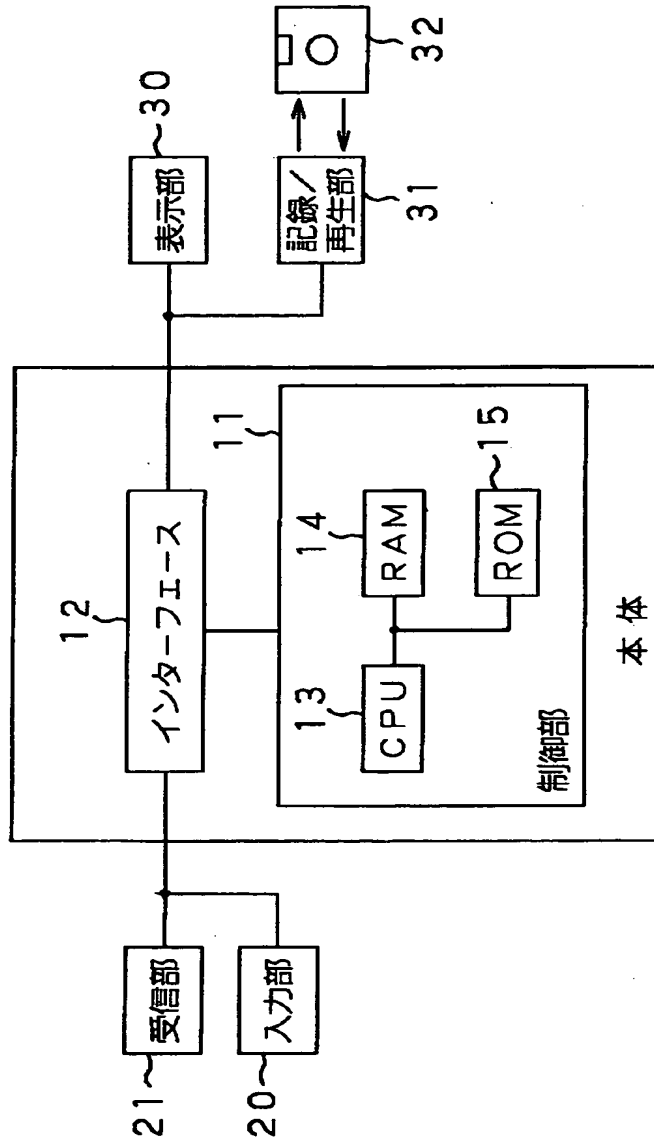
関連度による文書分類のフローチャートである。

【符号の説明】

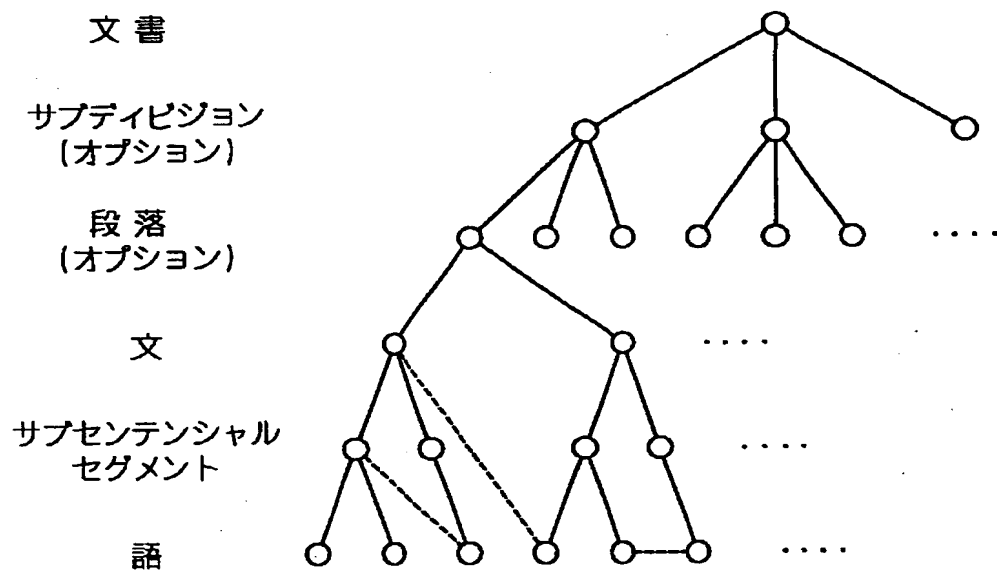
10 本体、11 制御部、12 インターフェース、13 CPU、20  
入力部、21 受信部、30 表示部、31 記録／再生部

【書類名】 図面

【図 1】



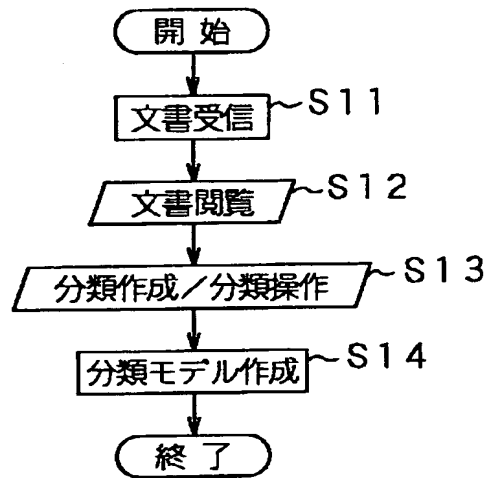
【図2】



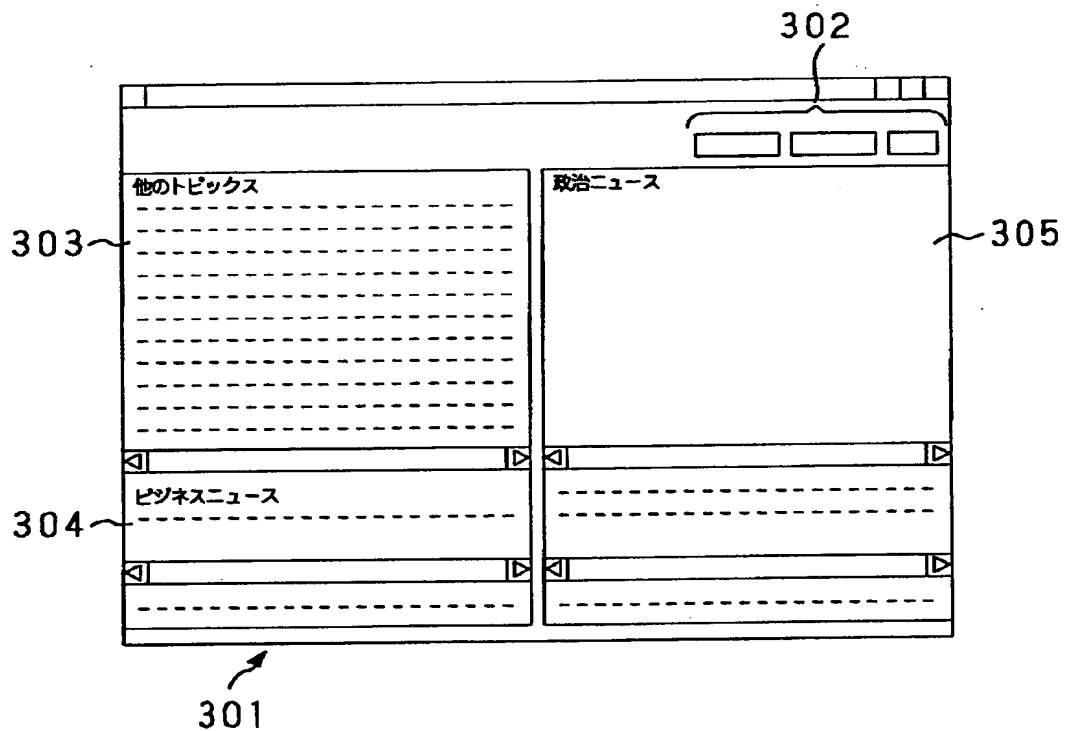




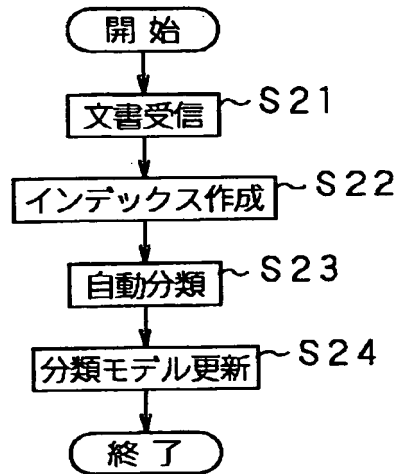
【図 4】



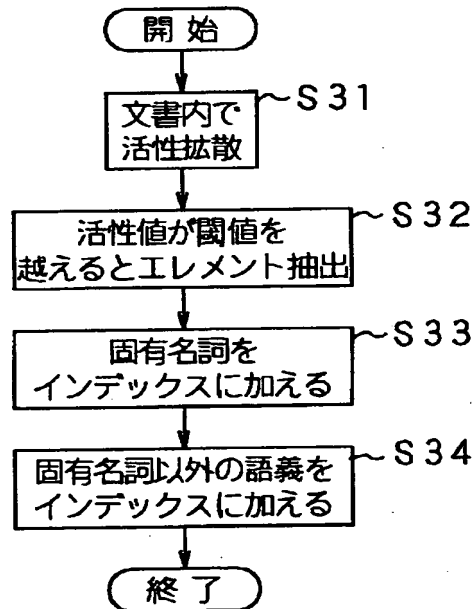
【図 5】



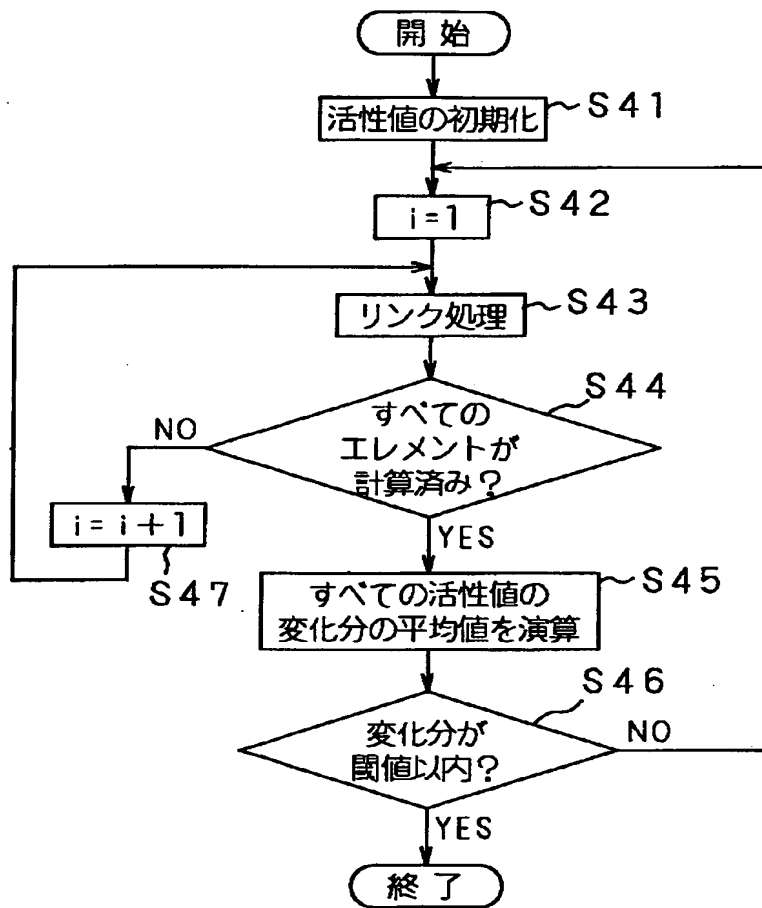
【図 6】



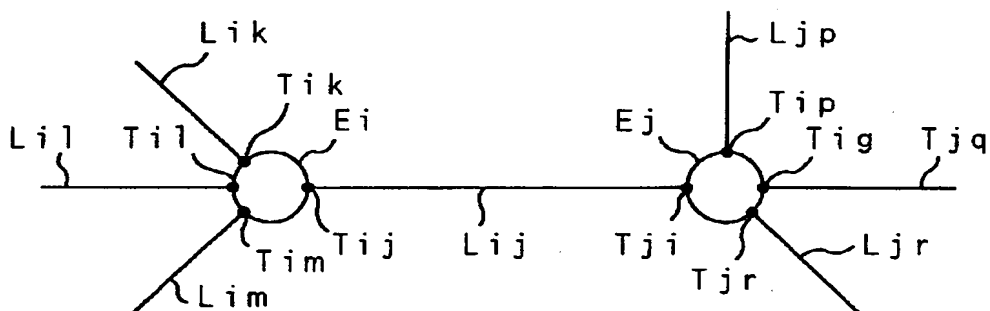
【図 7】



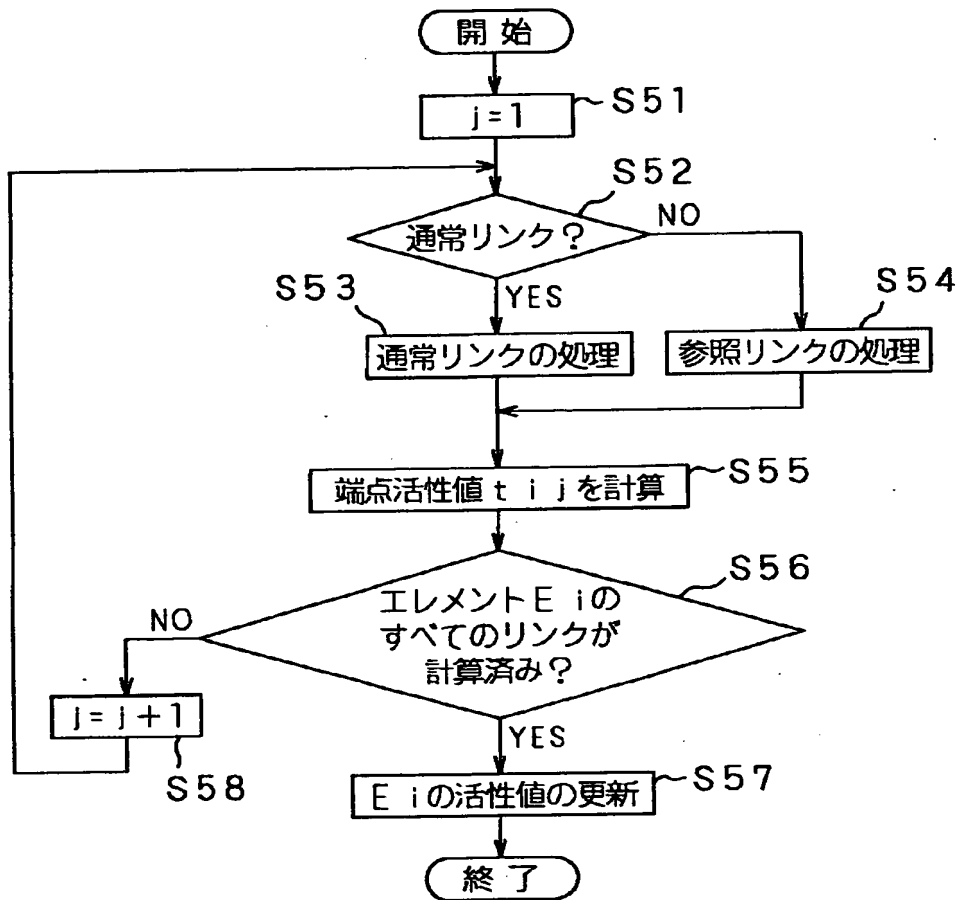
【図 8】



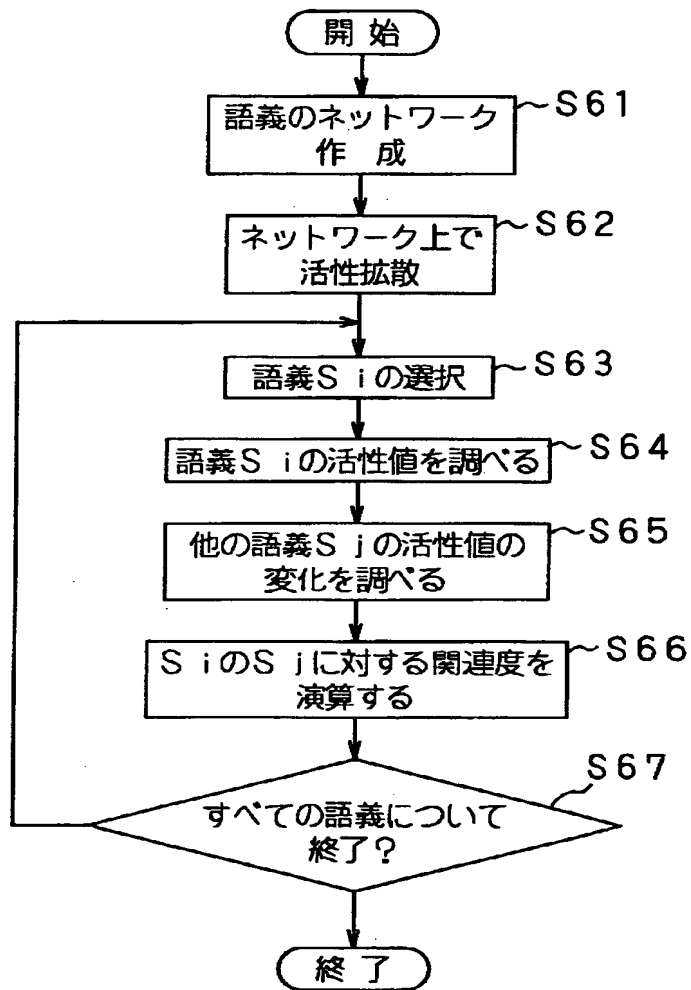
【図 9】



【図 10】



【図 1 1】



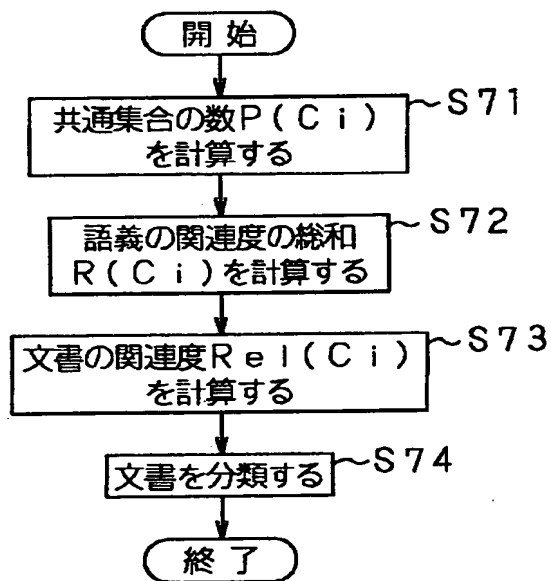
【図 1 2】

	コンピュータ	テレビ	
コンピュータ		0.55	
テレビ	0.55		
VTR	0.25	0.60	

【図 13】

分類項目	スポーツ	社会	コンピュータ	植物	美術	イベント
固有名詞	A 氏	B 社	C 社 G 社	D 種	E 氏	F 氏
語義	野球 グラウンド	労働 雇用	モバイル	桜1 (11111) オレンジ1	桜2 (11112) オレンジ2	桜3 (11113)
文書 アドレス	12345	22222	33333	44444	55555	66666

【図 14】



【書類名】 要約書

【要約】

【課題】 語義の関連度に基づいて文書を分類する。

【解決手段】 複数のエレメントから構成され、タグ付けによる内部構造を有する文書进行处理する文書処理装置であって、受信部 2 1 で受信した複数の文書をたとえば本体の制御部 1 1 の RAM 1 4 に記憶し、CPU 1 3 はたとえば ROM に記録された手順にしたがって、文書の特徴を表す特徴情報を抽出し、文書を分類する分類モデルを構成する複数の分類項目について、抽出した文書の特徴情報と上記分類項目ごとの特徴情報との関連度に応じて、各文書を分類項目に分類し、その結果を表示部 3 0 に表示する。

【選択図】 図 1



認定・付加情報

特許出願の番号	平成11年 特許願 第013307号
受付番号	59900049774
書類名	特許願
担当官	第七担当上席 0096
作成日	平成11年 1月25日

<認定情報・付加情報>

【提出日】	平成11年 1月21日
-------	-------------

出 願 人 履 歴 情 報

識別番号

[000002185]

1. 変更年月日	1990年 8月30日
[変更理由]	新規登録
住 所	東京都品川区北品川6丁目7番35号
氏 名	ソニー株式会社

**This Page is Inserted by IFW Indexing and Scanning  
Operations and is not part of the Official Record**

**BEST AVAILABLE IMAGES**

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☒ **BLACK BORDERS**
- ☐ **IMAGE CUT OFF AT TOP, BOTTOM OR SIDES**
- ☐ **FADED TEXT OR DRAWING**
- ☐ **BLURRED OR ILLEGIBLE TEXT OR DRAWING**
- ☐ **SKEWED/SLANTED IMAGES**
- ☐ **COLOR OR BLACK AND WHITE PHOTOGRAPHS**
- ☐ **GRAY SCALE DOCUMENTS**
- ☐ **LINES OR MARKS ON ORIGINAL DOCUMENT**
- ☐ **REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY**
- ☐ **OTHER:** \_\_\_\_\_

**IMAGES ARE BEST AVAILABLE COPY.**

**As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.**

**This Page Blank (uspto)**